DOCUMENT RESUME

ED 123 256 TM 005 315

AUTHOR Rubin, Rosalyn A.; Krus, Patricia H.
TITLE Multivariate Discriminant Function Predictions Employing Selected Missing Data Solutions.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C. Div. of Research.
PUB DATE [Apr 76]
GRANT OEG-32-33-0402-6021
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Comparative Analysis; *Data Analysis; *Discriminant Analysis; Early Childhood Education; Educable Mentally Handicapped; Longitudinal Studies; Measurement Techniques; *Prediction; Predictor Variables; *Research Problems; Statistical Analysis; *Student Placement
IDENTIFIERS *Missing Data

ABSTRACT

With a sample of 531 children measured on 79 variables, a series of discriminant function predictions of school placement were computed to compare the results of selected alternative procedures for treating missing data in multivariate data analyses. Procedures included elimination of cases with missing data as well as methods for substituting estimated values. While all procedures had some drawbacks, replacement of missing data with randomly drawn values from the available pool of study data resulted in least reduction in the multiple correlation while still maintaining a suitable sample size. Advantages and disadvantages of alternative procedures were discussed. (Author/DEP)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED123256

# MULTIVARIATE DISCRIMINANT FUNCTION PREDICTIONS EMPLOYING SELECTED MISSING DATA SOLUTIONS

Rosalyn A. Rubin

University of Minnesota

Patricia H. Krus

Arizona State University

PERMISSION TO REPRODUCE THIS MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY

ROSALYN A. RUBIN

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

U.S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 315

Paper presented at Session 8.14 of the Annual Meeting of the American Educational Research Association in San Francisco, April 1976

Abstract

MULTIVARIATE DISCRIMINANT FUNCTION PREDICTIONS

EMPLOYING SELECTED MISSING DATA SOLUTIONS

Rosalyn A. Rubin
University of Minnesota

Patricia H. Krus
Arizona State University

With a sample of 531 children measured on 79 variables, the present study computed a series of discriminant function predictions of school placement to compare the results of selected alternative procedures for treating missing data in multivariate data analyses. Procedures included elimination of cases with missing data as well as methods for substituting estimated values. While all procedures had some drawbacks, replacement of missing data with randomly drawn values from the available pool of study data resulted in the least reduction in the multiple correlation while still maintaining a suitable sample size. Advantages and disadvantages of alternative procedures were discussed.

# MULTIVARIATE DISCRIMINANT FUNCTION PREDICTIONS EMPLOYING SELECTED MISSING DATA SOLUTIONS[1]

Rosalyn A. Rubin
University of Minnesota

Patricia H. Krus
Arizona State University

The purpose of the present study was to determine optimal procedures for treating missing data in multivariate data analyses. Comparisons of alternative procedures were based on an extensive set of real data with the predictor variables consisting of early medical, socioeconomic and psychological data while the criterion to be predicted was school placement in special classes for the educable mentally retarded.

When multivariate data analysis techniques are employed on a large pool of subjects, it is not unusual to find that one or more of the observations in the variable data matrix are not obtainable. This is a problem of particular significance in conducting longitudinal investigations in which the same subjects are administered repeated measures over an extended period of time, and not all subjects are available at the appropriate time for all of the requisite observations. The more extensive the investi-

---

[1] The research reported herein was performed pursuant to a grant (OEG-32-33-0402-6021) from the Division of Research, Bureau of Education for the Handicapped, U.S. Office of Education, Department of Health, Education and Welfare. Data were also made available through the cooperation of the Minnesota section of the Collaborative Project supported by the National Institute of Neurological Diseases and Stroke (Public Health Service grant PH-43-68-9).

The authors wish to thank Ruoh-Yun Hsieh for her contributions in programming and computer analysis of study data.

gation in terms of number of subjects and amount of information collected, the greater the opportunity for missing data to become a problem.

Various techniques for processing missing data have been suggested for use in those instances where incomplete data are unavoidable (Afifi & Elashoff, 1966, 1969; Guertin, 1968). In a literature review covering the past 40 years, we found that 75% of the available research addressed to the problem of missing data involved mathematical proofs and often no data. Only 10% of the available research papers were based on real, nonsimulated data, and only one paper employed psychological measures, thus illustrating the paucity of research on the topic of missing data based on real data examples on human subjects.

The present study provides a comparison of four major techniques for dealing with missing data applied to a set of real data gathered as part of the Educational Follow-Up Project, a longitudinal investigation of the educational and behavioral sequelae of a number of perinatal and early childhood conditions. A major portion of the data analysis activities of this longitudinal study involves the use of multivariate regression techniques in the prediction of educational and behavioral outcomes from early family and medical data. Since all observations or their estimates are required for most correlation-based multivariate statistical analysis (e.g., linear multiple regression and factor analytic techniques), it becomes necessary to identify the optimal means for treating missing observations in the data matrix.

## Data Source

Subjects were drawn from among the 1559 participants in the Educational Follow-Up Project (EFUP), a prospective study of long-term educational and behavioral outcomes associated with conditions of pregnancy, delivery,

infancy and early childhood (Balow, et al., 1969). The total EFUP population is essentially normal on measures of social class, intelligence, and medical conditions at the time of birth.

In the present investigation children from this basic subject pool who had been placed in elementary school classes for the educable mentally retarded (EMR) were compared with children who had no record of school-identified problems of learning or behavior. There were 30 children in the EMR group and 501 in the "no-problem" regular class control group.

Method and techniques

Observations were available for all study subjects on the criterion variable, school placement, as well as on 8 (10%) of the 78 perinatal and infant predictor variables. Less than 1% of the observations were missing on 38 (49%) of the variables, and in no instance were more than 10% of the observations missing on any single variable included in the analysis.

Point-biserial correlations predicting the criterion variable of school placement from the set of 78 predictor variables were computed under each of the following conditions: (a) Listwise Deletion--all subjects with any missing data were eliminated from the analysis; (b) Pairwise Deletion--subjects were eliminated only from computations involving specific variables on which they had missing data. This process involved computation of covariances between all pairs of variables, each time using only those subjects with recorded values for both variables. The covariances thus obtained were used in constructing the system of normal equations. One problem with this method is that the subjects retained in the analysis differ for each variable pair correlation. Therefore the sample size cannot be

specified; (c) Mean Replacement--all missing data were replaced by the mean value of the available observations for that variable for subjects in the study sample; (d) Random Replacement--all missing data were replaced by randomly drawn values from the available pool of observations for that variable on study subjects (this should be understood as different from replacement with numbers from a table of random numbers).

Deletion of all variables with missing observations rather than deletion of subjects with missing data was not attempted since this procedure would have eliminated 90% of the predictor variables.

Three of the procedures for treating missing data were then compared in a discriminant function prediction of group membership in special (EMR) vs. regular class. It was not possible to compute the accuracy of classification under Pairwise Deletion because of the previously noted variations in sample size for each variable pair correlation.

## Results

---

Table 1 about here

---

When estimated values were used to replace missing values, sample size doubled for the EMR group (from N of 15 to 30) and increased by 53% for the control group (from 328 to 501). Relatively small differences in prediction, accounting for no more than 10% of the total variance, were found among the four procedures. Pairwise Deletion of subjects with missing data appeared to be the least desirable because it was one of the poor predictors of the criterion and because of the fluctuation of sample size during the course of the analysis. Listwise Deletion of subjects

with missing data enables clear identification of the sample and also yielded the highest predictive correlation with the criterion. However, this procedure resulted in the loss of 35% of the available subject pool for whom extensive data had already been collected.

---

Table 2 about here

---

Accuracy of classification of control and EMR subjects under three of the missing data conditions was computed using stepwise discriminant analysis (see Table 2).

Replacement of missing data with estimated values enables retention of a much higher proportion of cases but results in some reduction in accuracy of prediction. In comparing two methods for estimating the missing values, Mean Replacement vs. Random Replacement, it was found that Mean Replacement led to a 20% reduction in the accuracy of classifying members of the experimental group. However, Random Replacement resulted in a 3% reduction in accuracy of classifying the control group.

## Conclusions

Whenever a large pool of subjects is measured over many variables, particularly in situations of repeated testing, it is not unusual to find that one or more of the observations are missing with no opportunity of recouping the missing data. Most correlational techniques require all observations or their estimates. Thus it becomes necessary to identify the optimal means for treating missing observations.

The goals of any particular investigation would help to determine which procedure would be most appropriate. In circumstances such as the

present study wherein the loss of valuable data is deemed of greater practical significance than is a limited reduction in predictive accuracy, it would therefore appear advantageous to substitute randomly drawn values from the study pool as estimates of the missing observations.

While there is no completely satisfactory resolution of the problem inherent in the analysis of data with missing observations, it is possible to examine differences in prediction between alternative procedures. The systematic comparison of the outcomes of procedures applied in the present study provides information to other investigators which can serve as a basis for decision making in situations involving incomplete data matrices.

## References

Afifi, A. A., & Elashoff, R. M. Missing observations in multivariate statistics, I: Review of the literature. Journal of the American Statistical Association, 1966, 61, 595-604.

Afifi, A. A., & Elashoff, R. M. Missing observations in multivariate statistics, III: Large sample analysis of simple linear regression. Journal of the American Statistical Association, 1969, 64, 337-358.

Balow, B., Anderson, J., Reynolds, M., & Rubin, R. Educational and behavioral sequelae of prenatal and perinatal conditions. (USOE, BEH, Project No. 6-1176, Interim Report No. 3) Department of Special Education, University of Minnesota, September 1969.

Guertin, W. H. Comparison of three methods of handling missing observations. Psychological Reports, 1968, 22, 896.

Table 1

Point-Biserial Correlation Coefficients Obtained
Under Various Missing Data Solutions
(Based On 78 Predictor Variables)

| | EMR N | Control N | R |
|---|---|---|---|
| Listwise Deletion | 15 | 328 | .55 |
| Pairwise Deletion | unknown | unknown | .47 |
| Replacement with Mean Values | 30 | 501 | .45 |
| Replacement with Random Values from Data Pool | 30 | 501 | .50 |

Table 2

Accuracy of Prediction of Classification Under Missing Data and Estimated Values Added Procedures

| Actual Criterion Group | Group Membership Based on Discriminant Function Predictions: Frequency EMR | Control | Correct Group Classifications (%) | Total Correct (%) |
|---|---|---|---|---|
| Missing Data-Listwise Deletion | | | | |
| EMR (N=15) | 12 | 3 | 80.0 | |
| Control (N=328) | 23 | 305 | 93.0 | 92.4 |
| Mean Replacement | | | | |
| EMR (N=30) | 19 | 11 | 63.3 | |
| Control (N=501) | 42 | 459 | 91.6 | 90.2 |
| Random Replacement | | | | |
| EMR (N=30) | 25 | 5 | 83.3 | |
| Control (N=501) | 58 | 443 | 88.4 | 88.1 |